SynthSys
Centre for Synthetic and Systems Biology

Innovating research, policy and education in synthetic and systems biology

Balancing shared infrastructure and domain-specific tools in data management

Andrew Millar, Eilidh Troup, Tomasz Zieliński

@A_J_Millar

EPSRC
Engineering and Physical Sciences Research Council

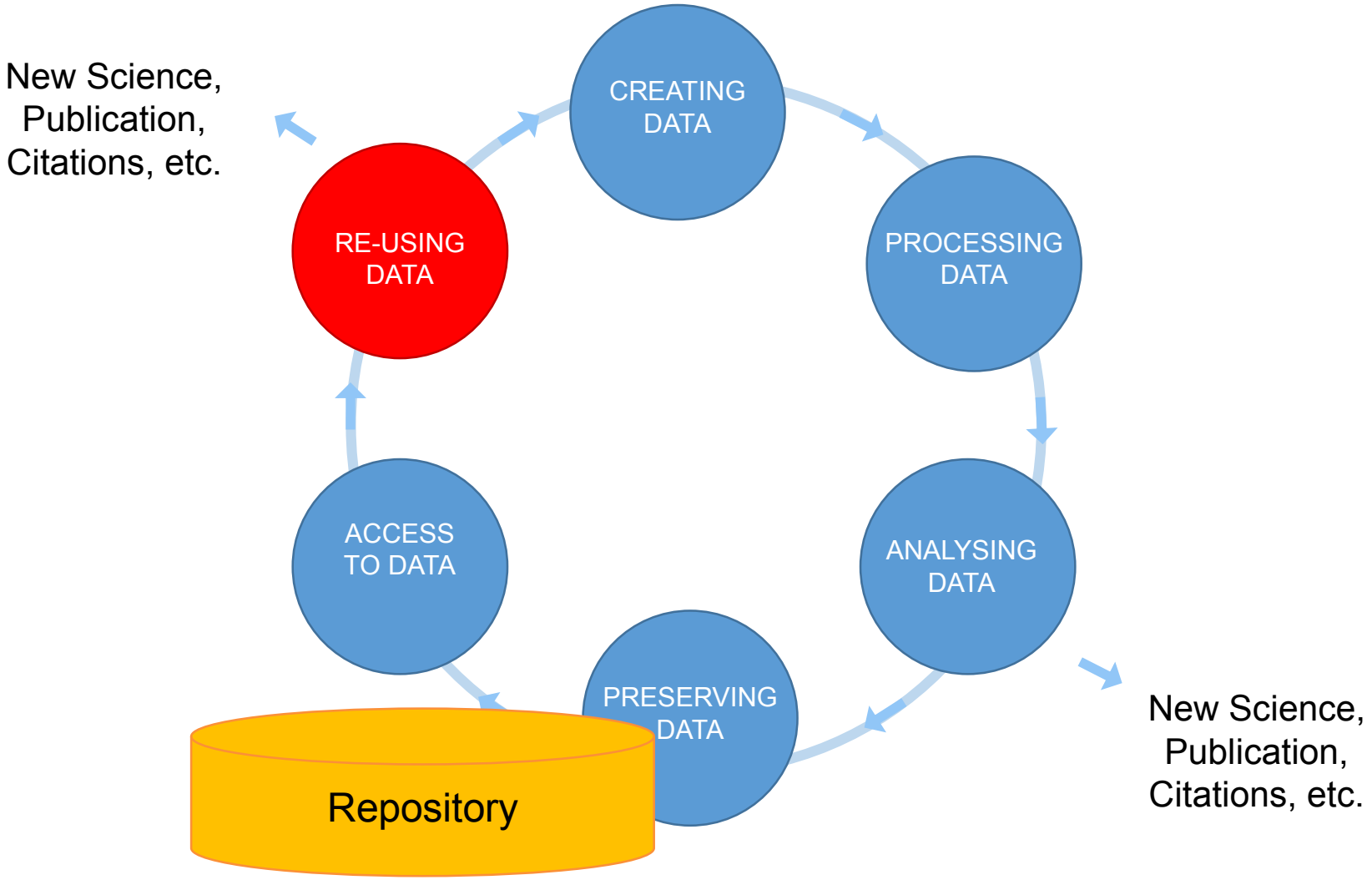
BBSRC
bioscience for the future

MRC Medical Research Council

THE UNIVERSITY of EDINBURGH

UK Centre for Mammalian Synthetic Biology

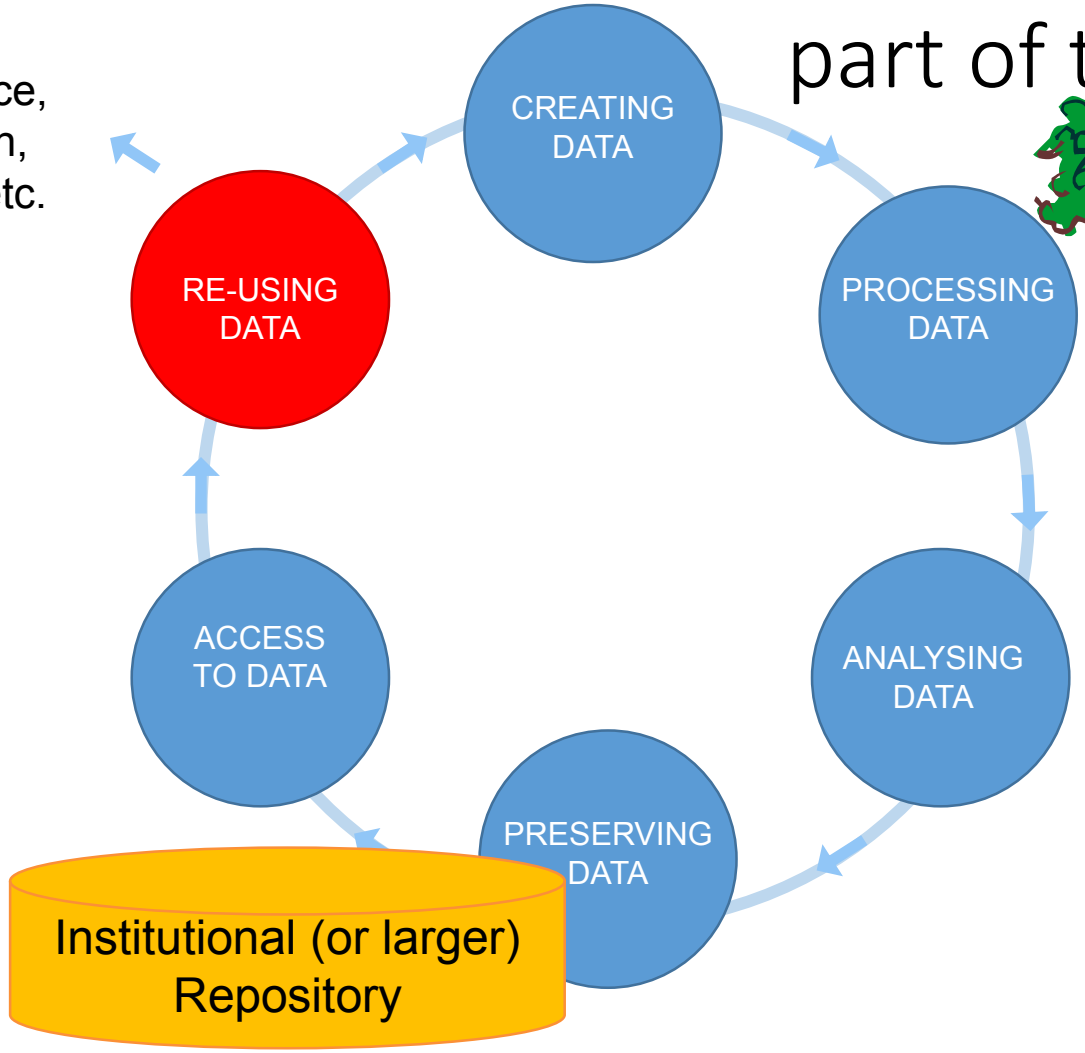
Data life-cycle



Take-home message

Data management as part of the workflow

New Science, Publication, Citations, etc.

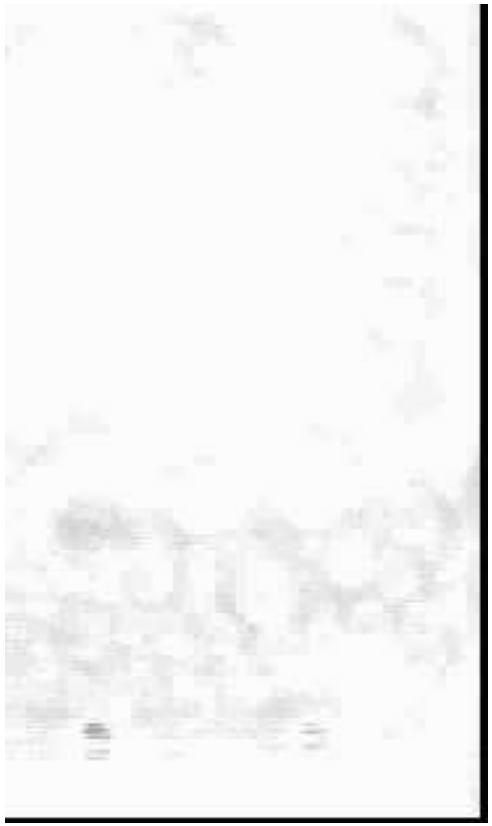


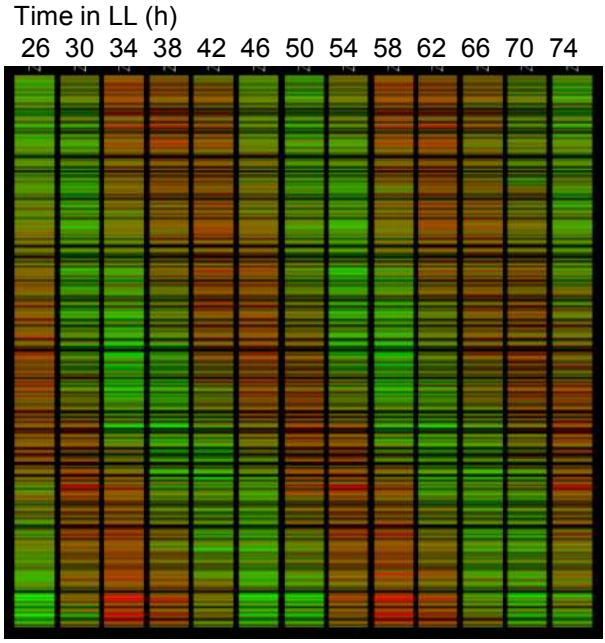
Intermediate Systems

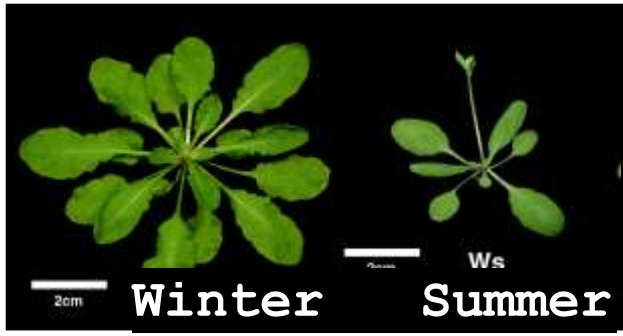
FAIRDOM

Successful RDM needs Researchers AND Institutions, linked through intermediate RDM systems.

The Biological Clock



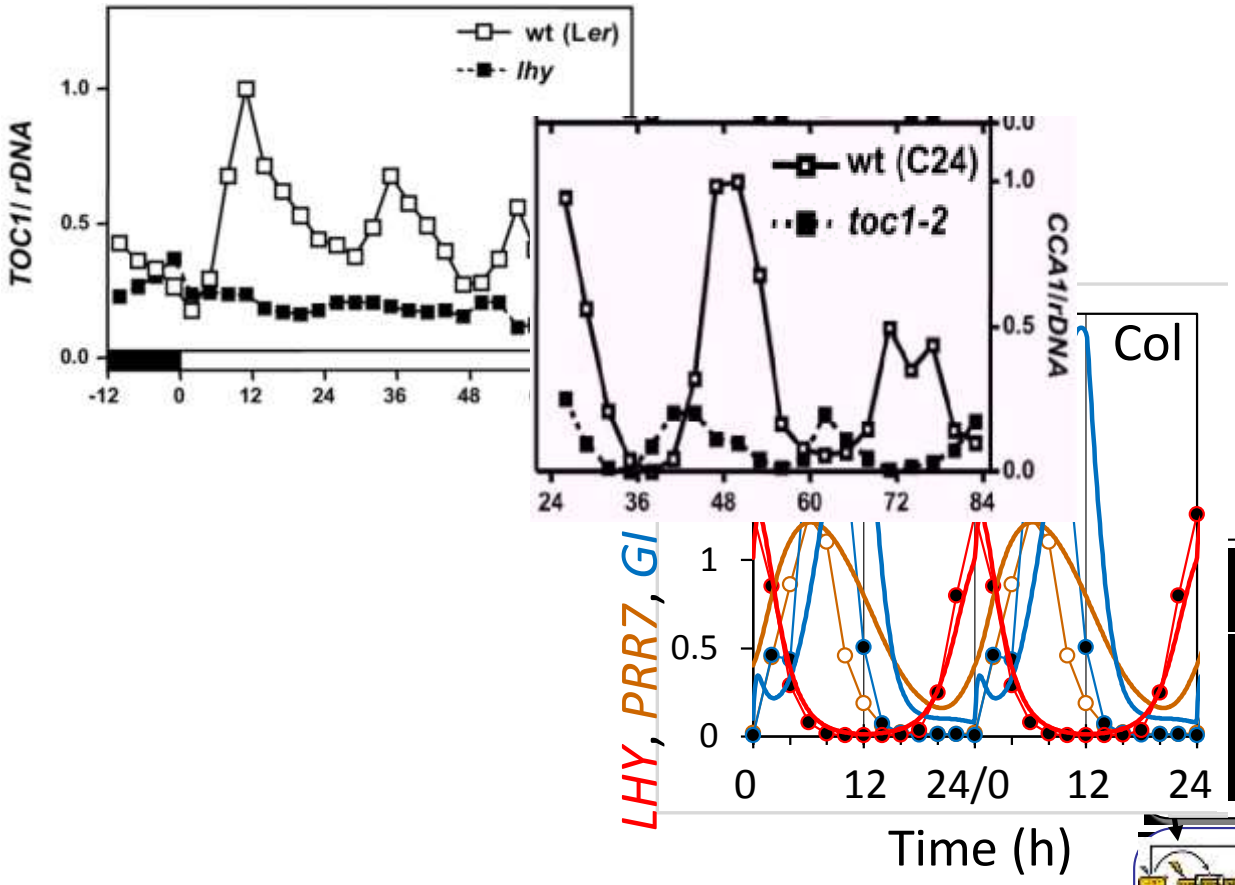


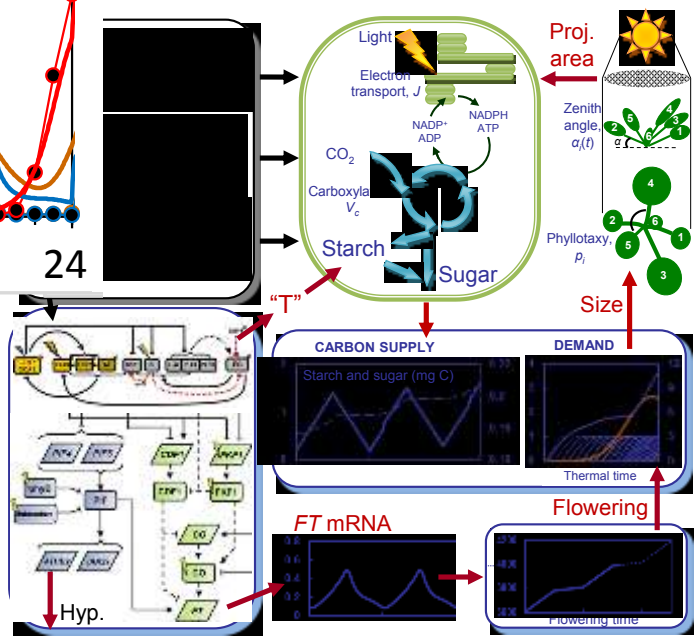


- Controls >30% of genes, elongation, flowering.
- Models connect gene level to whole plant growth

Dowson-Day *et al.* TPJ '99
Edwards *et al.* Plant Cell '06
Salazar *et al.* Cell '09

Variety of data from … graphs





- Big data? Variety!

Pokhilko *et al.* Mol. Syst. Biol.'12
Seaton *et al.* Mol. Syst. Biol.'15
Chew *et al.* bioRxiv '17

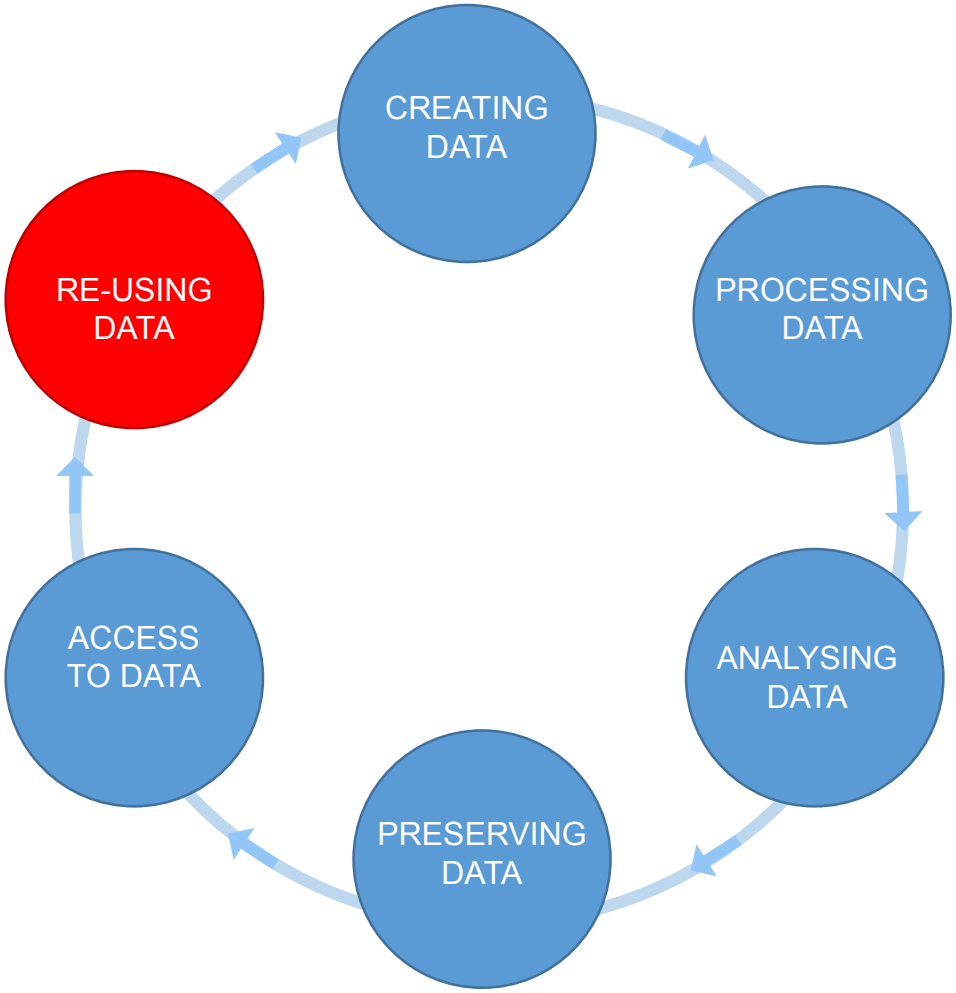
Variety of data from … graphs

Scraping PDFs …

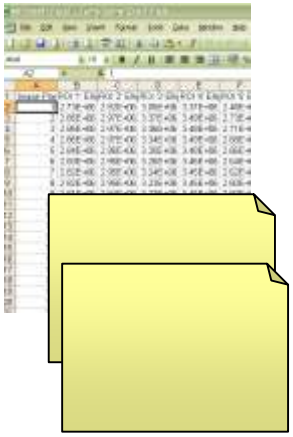
> …is a bit like cleaning drains with your teeth.
>
> It's slow, unpleasant,
>
> and you can't help but feel you're using
> the wrong tools for the job.
>
> (annapowellsmith, Scraperwiki blog, 2010)

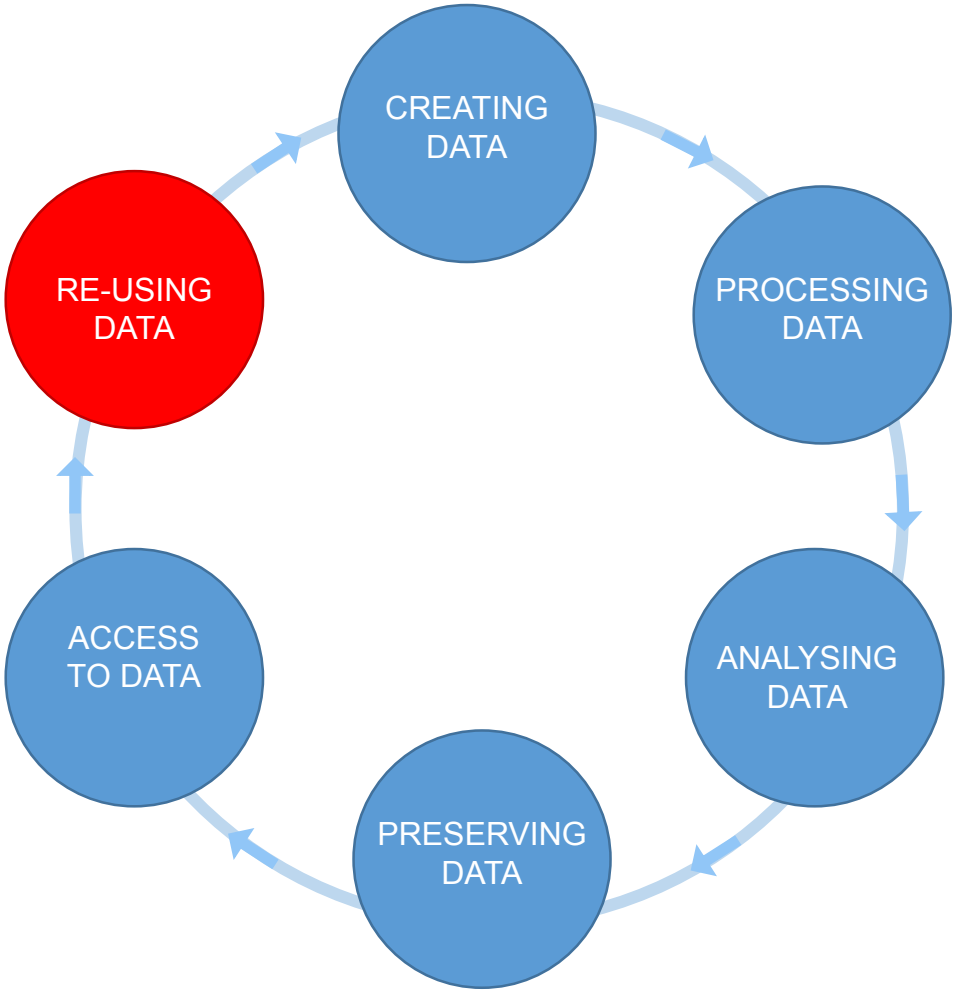
… a motivation for Open Research Data.

Data life-cycle

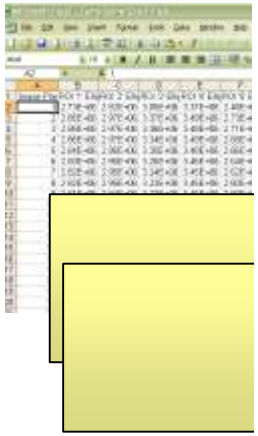


Data life-cycle





Reality of Data Sharing

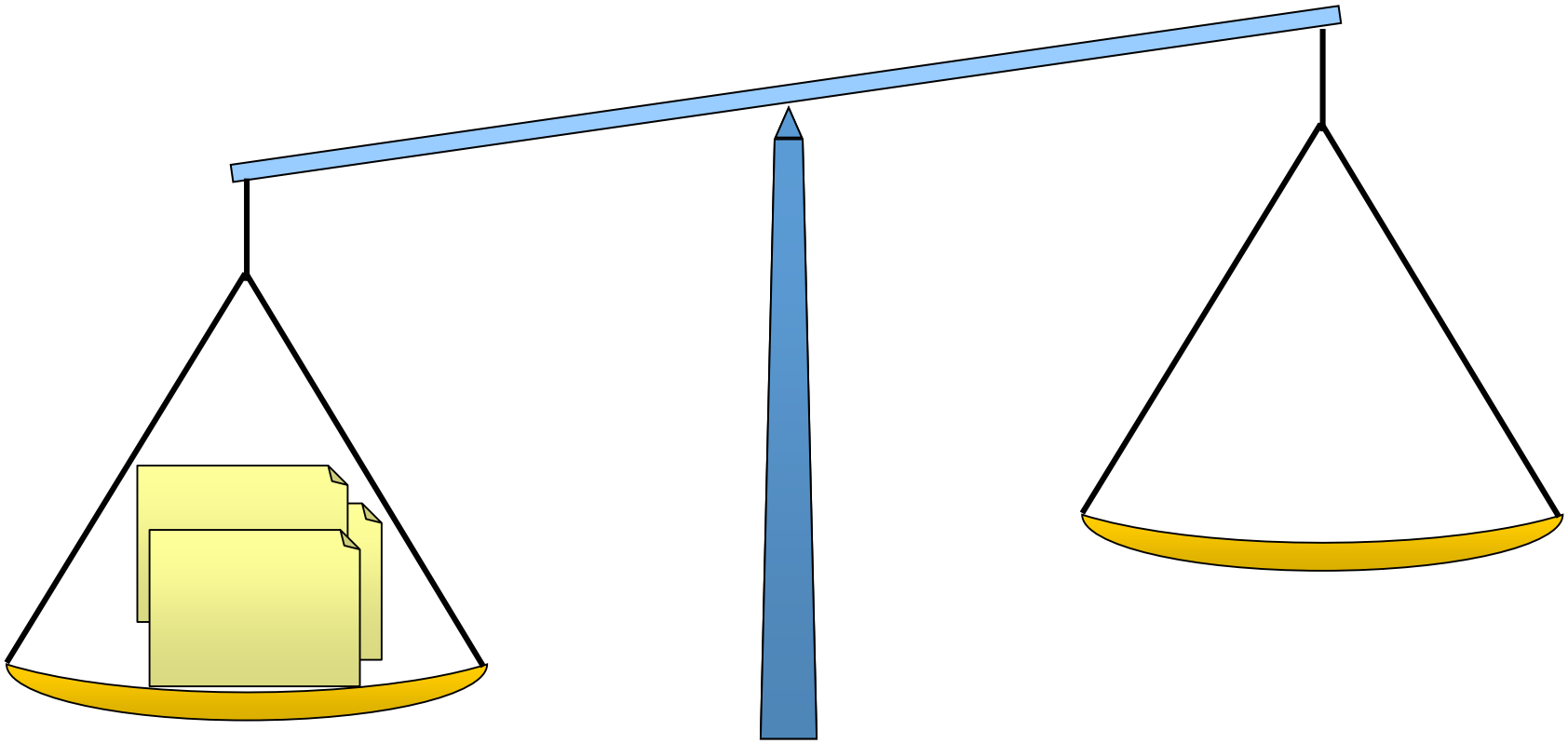


Data Creator:	Martin, Sarah F.; Noordally, Zeenat B.; van Ooijen, Gerben; Barrios-Llerena, Martin E.; Simpson, T. Ian; Millar, Andrew J.; Hindle, Matthew M.; Thierry Le Bihan
Date Available:	2014-06-24
Citation:	Martin, Sarah F.; Noordally, Zeenat B.; van Ooijen, Gerben; Barrios-Llerena, Martin E.; Simpson, T. Ian; Millar, Andrew J.; Hindle, Matthew M.; Thierry Le Bihan. (2014). The reduced kinome of Ostreococcus tauri: core eukaryotic signalling components in a tractable model species, [Dataset]. University of Edinburgh. SynthSys and School of Biological Sciences. http://dx.doi.org/10.7488/ds/72.
Dataset Description (abstract):	Background The current knowledge of eukaryote signalling originates from phenotypically diverse organisms. There is a pressing need to identify conserved signalling components among eukaryotes, which will lead to the transfer of knowledge across kingdoms. Two useful properties of a eukaryote model for signalling are (1) reduced signalling complexity, (2) conservation of signalling components. The alga Ostreococcus tauri is described as the smallest free-living eukaryote. With less than 8,000 genes, it represents a highly constrained genomic palette. Results Our survey revealed 133 protein kinases and 34 protein phosphatases (1.7% and 0.4% of the proteome). We conducted phospho-proteomic experiments and constructed domain structures and phylogenies for the catalytic protein-kinases. For each of the major kinases families we review the completeness and divergence of O. tauri representatives in comparison to the well-studied kinomes of the laboratory models Arabidopsis thaliana and Saccharomyces cerevisiae, and of Homo sapiens. Many kinase clades in O. tauri were reduced to a single member, in preference to the loss of family diversity, whereas TKL and ABC1 clades were expanded. We also identified kinases that have been lost in A. thaliana but retained in O. tauri. For three, contrasting eukaryotic pathways – TOR, MAPK, and the circadian clock – we established the subset of conserved components and demonstrate conserved sites of substrate phosphorylation and kinase motifs. Conclusions We conclude that O. tauri satisfies our two central requirements. Several of its kinases are more closely related to H. sapiens orthologs than S. cerevisiae is to H. sapiens. The greatly reduced kinome of O. tauri is therefore a suitable model for signalling in free-living eukaryotes.

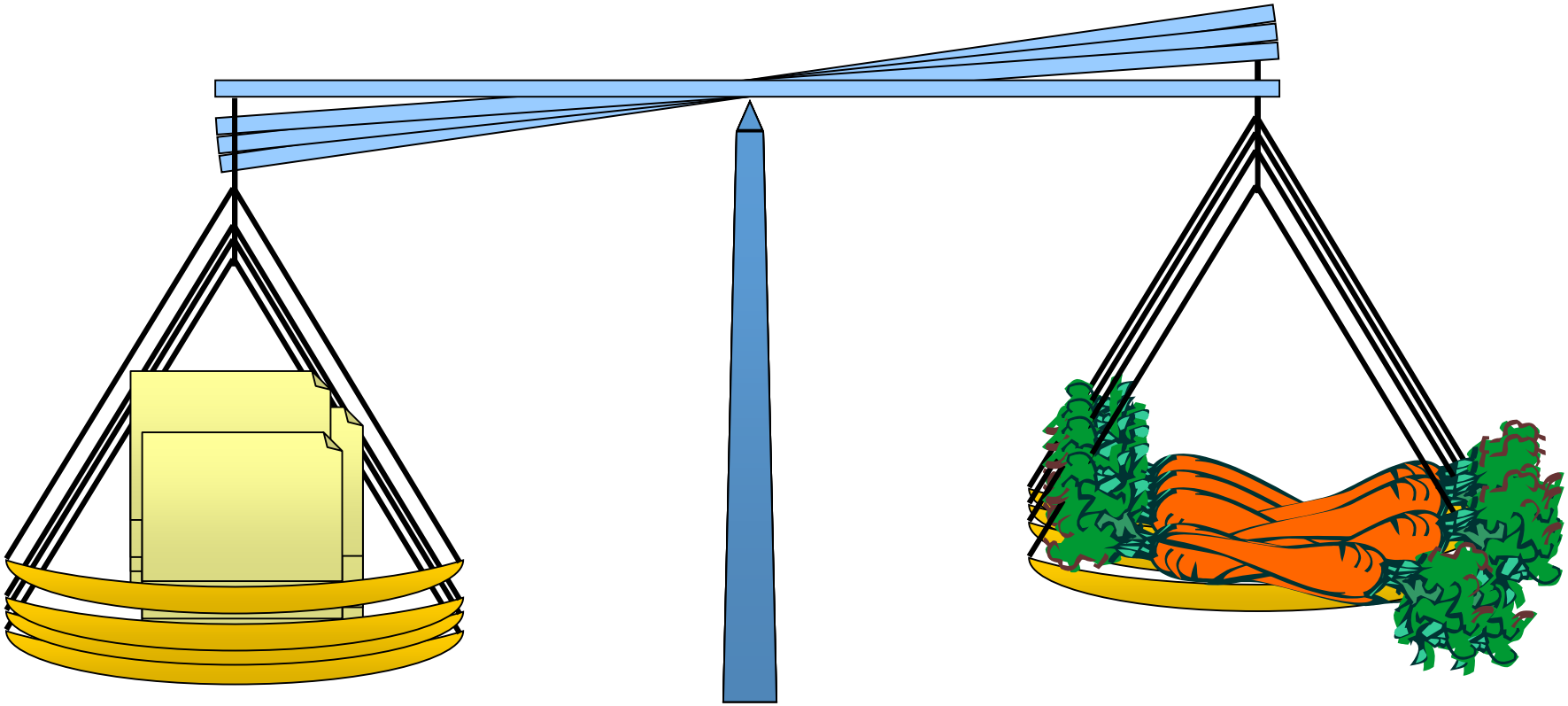
Reality of Data Sharing

Getting metadata

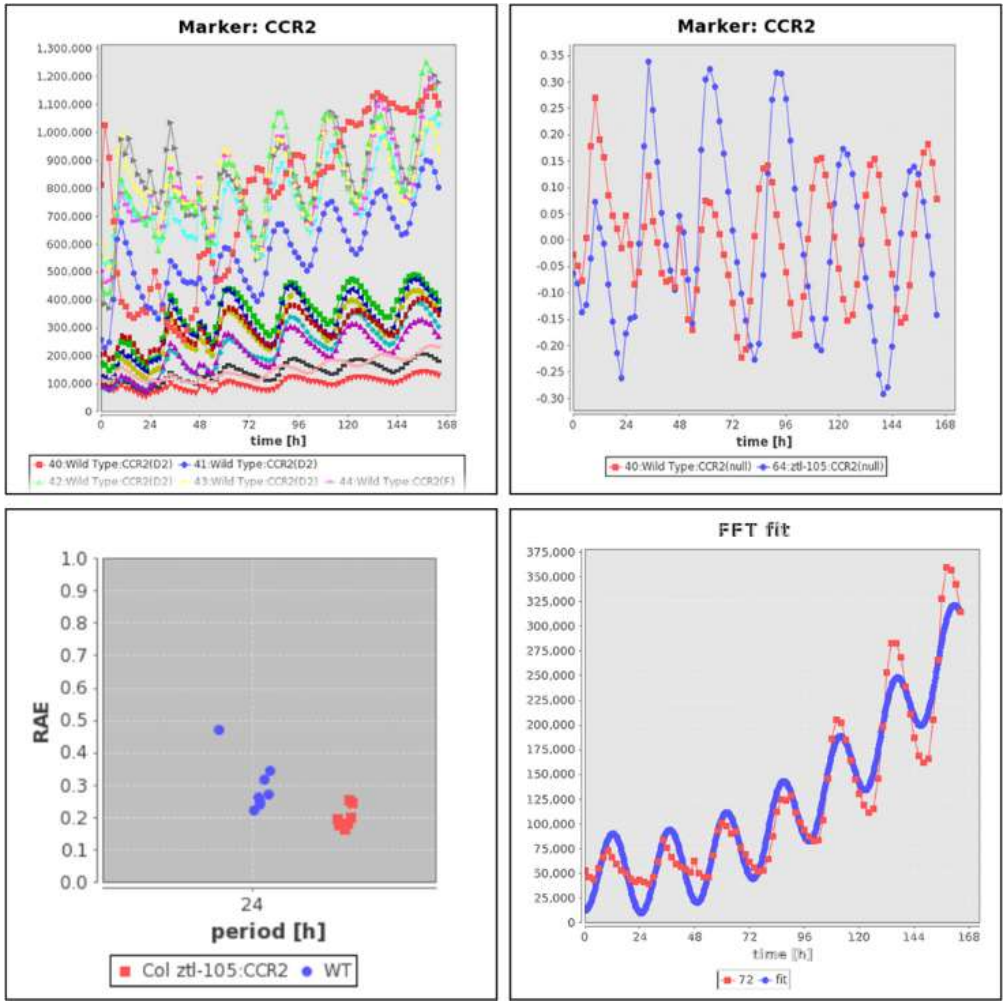
- No one likes describing data for repositories



Getting metadata - Tipping the balance



Getting metadata - Tipping the balance



BioDare:

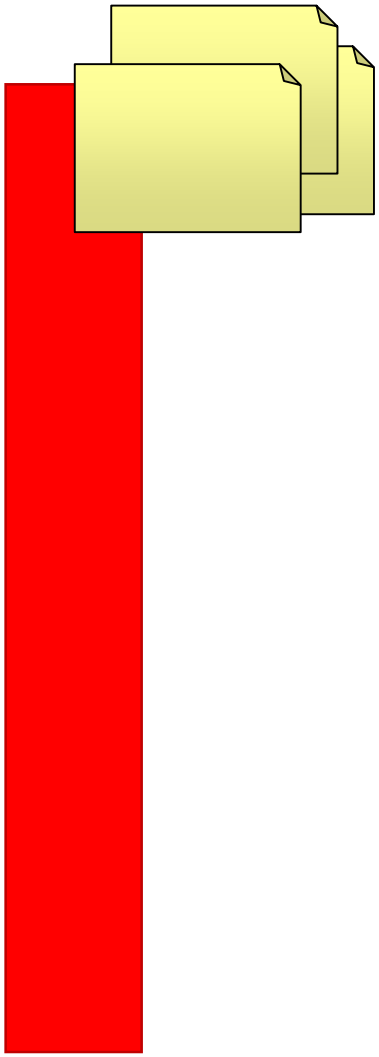
Biological rhythm data repository

- data visualization
- timeseries processing
- fast rhythm analysis
- good user interface

Immediate value for researchers.

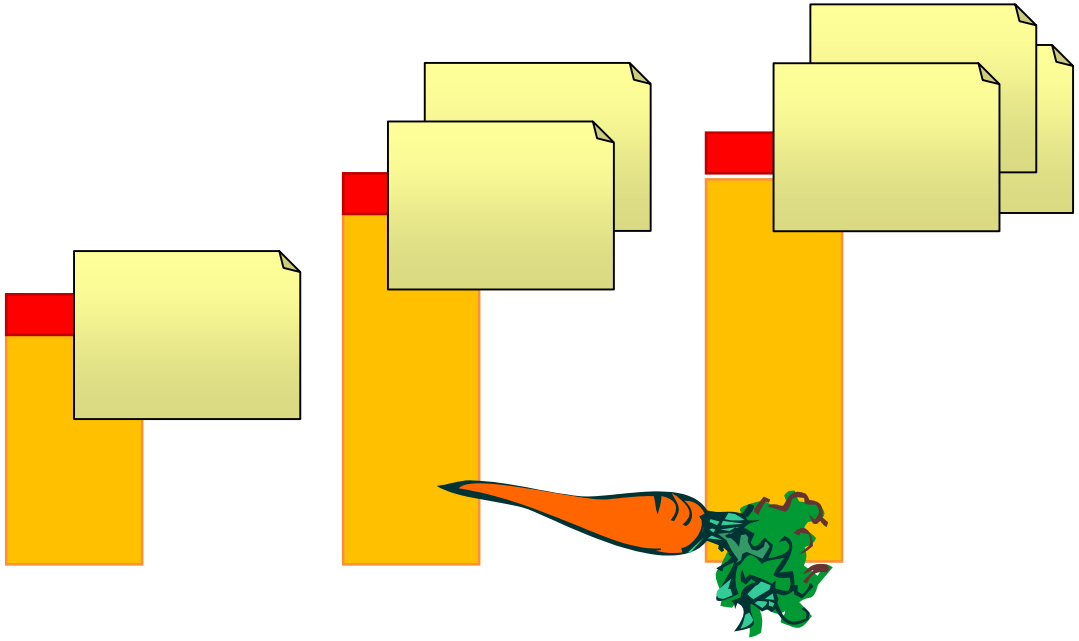
www.biodare2.ed.ac.uk

Lowering the barrier



- Also splitting the effort

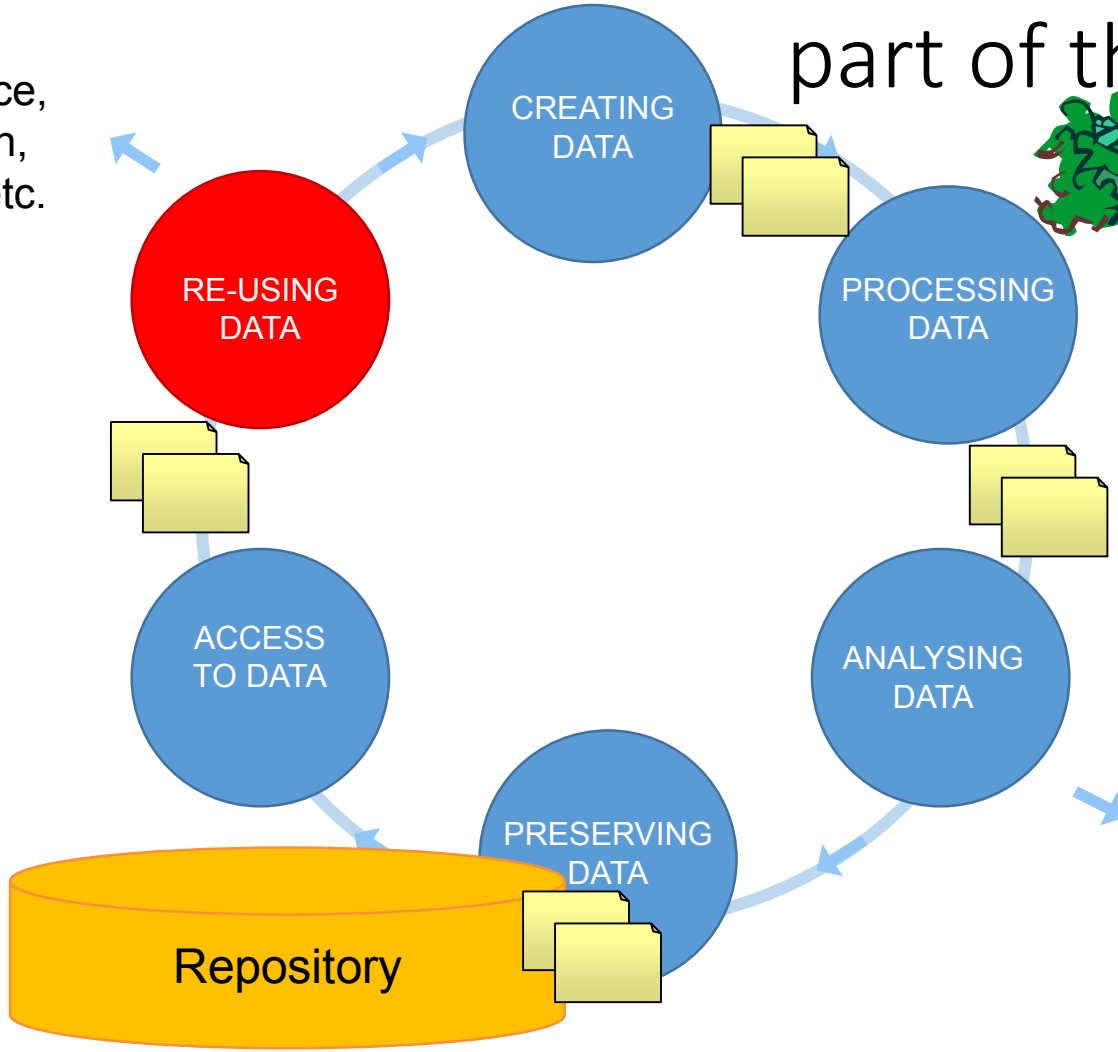
- splitting among people
- splitting the description



Data life-cycle

Data management as part of the workflow

New Science, Publication, Citations, etc.

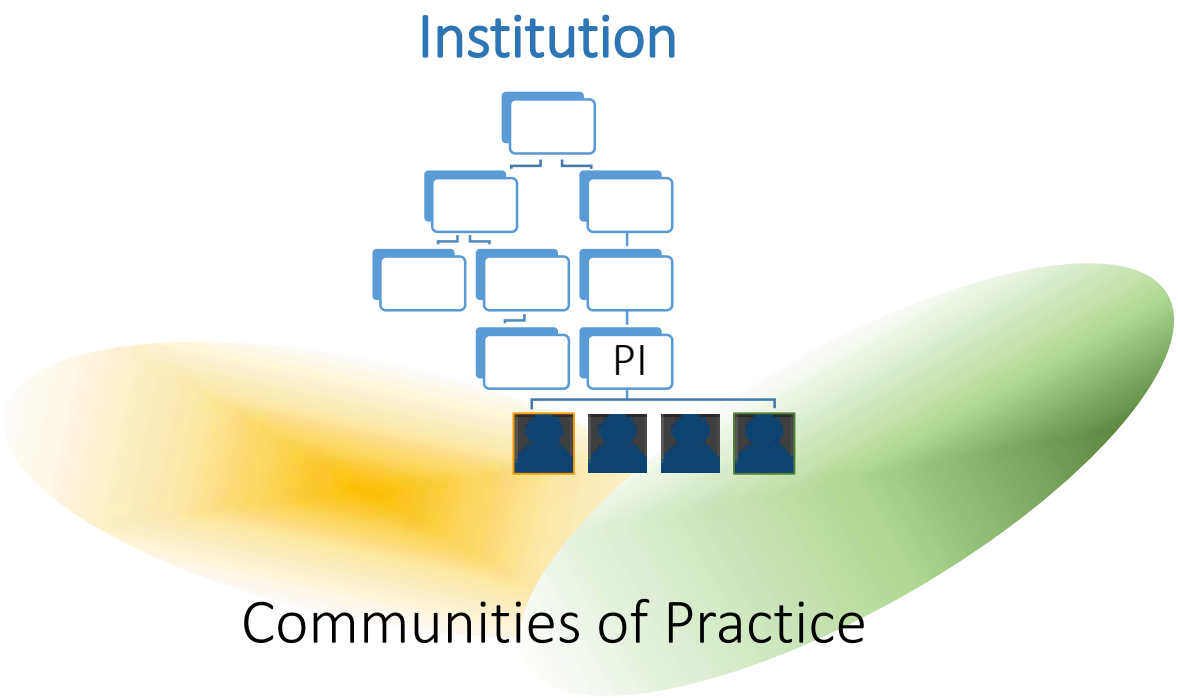


Concerns:

- durability
- support

New Science, Publication, Citations, etc.

Research organisation



Freeman and Millar, “Valuing the Project”, Public Money & Management, 2017

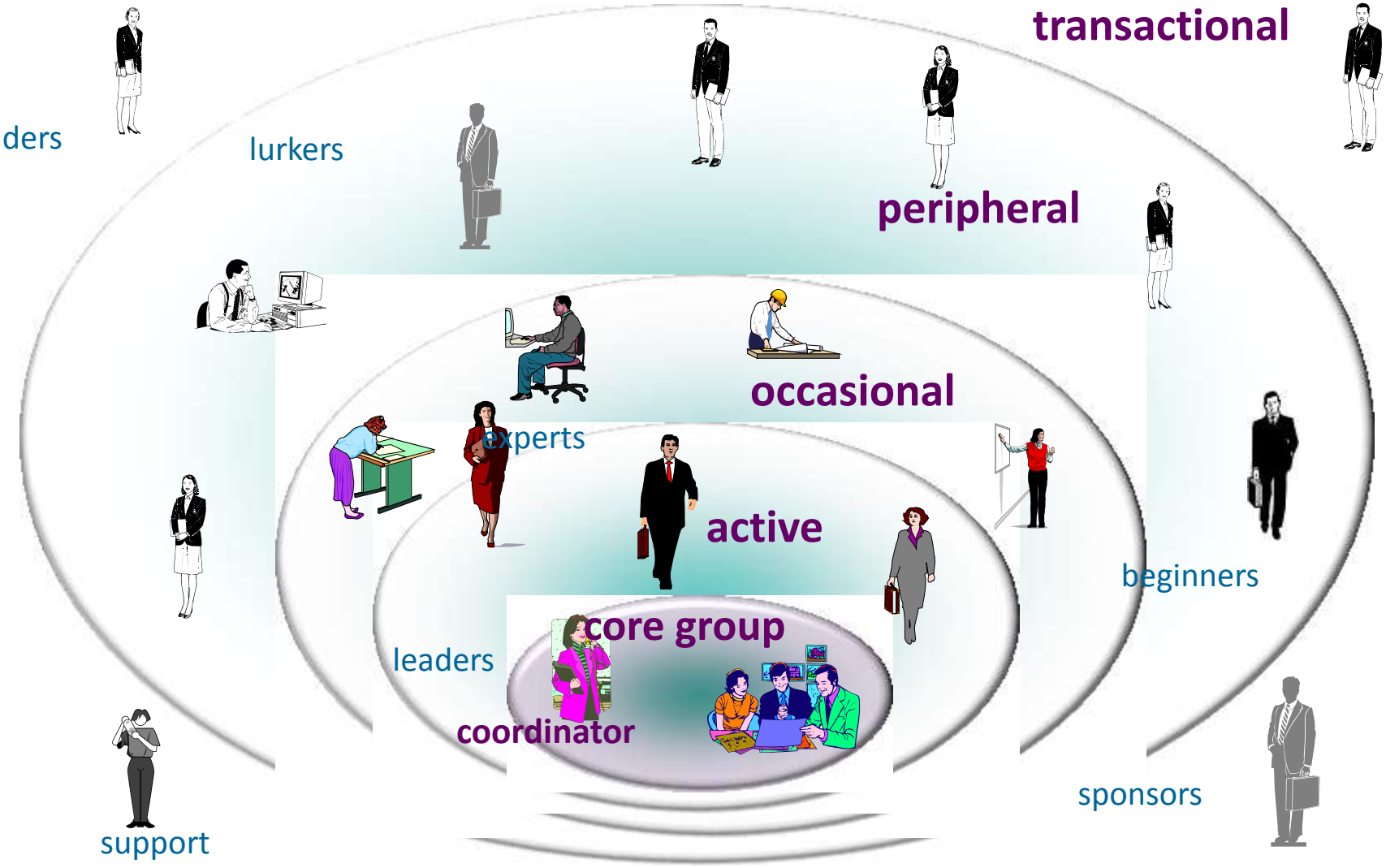
Communities of Practice (CoP)

Groups of people who share a concern, a set of problems, or a passion about a topic, and who **deepen their knowledge and expertise** by interacting on an ongoing basis.

(Etienne Wenger)

Levels of participation

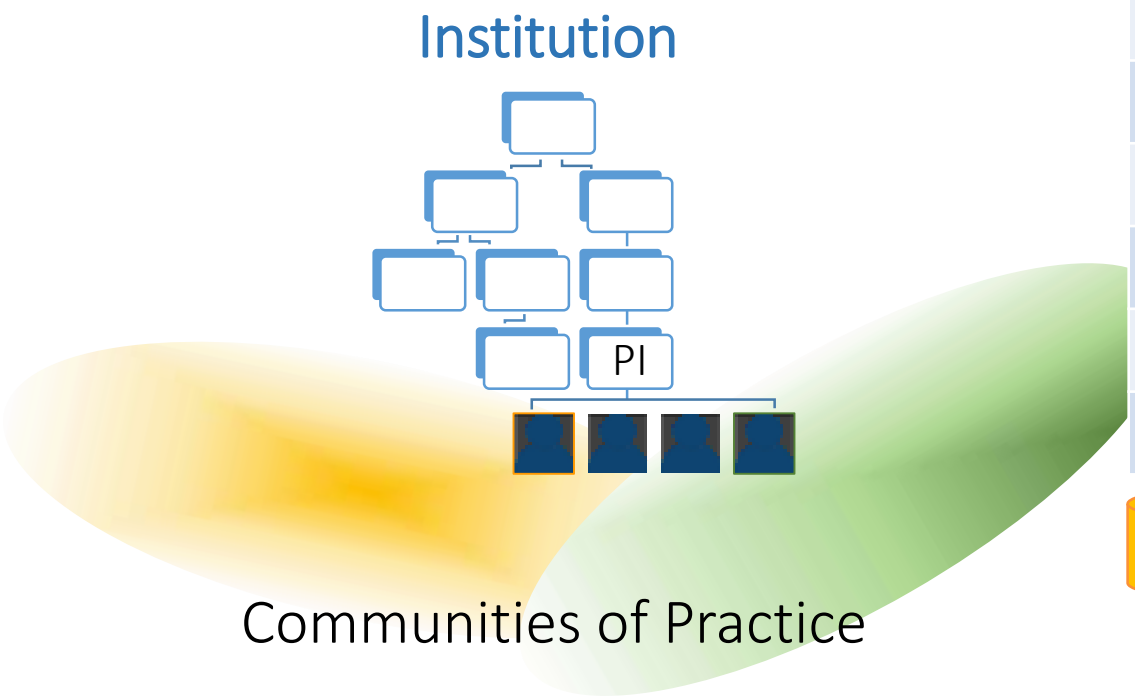
- a common picture

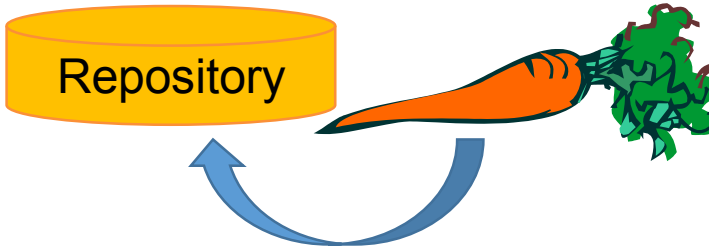


Etienne and Beverly Wenger-Trayner

Research organisation

Institution	Community
Hierarchy	Network
Rigid	Fluid
Competitive	Collaborative
Visible	Obscure
Persistent	Transient
Professional	Best efforts
Broad	(Very) specific





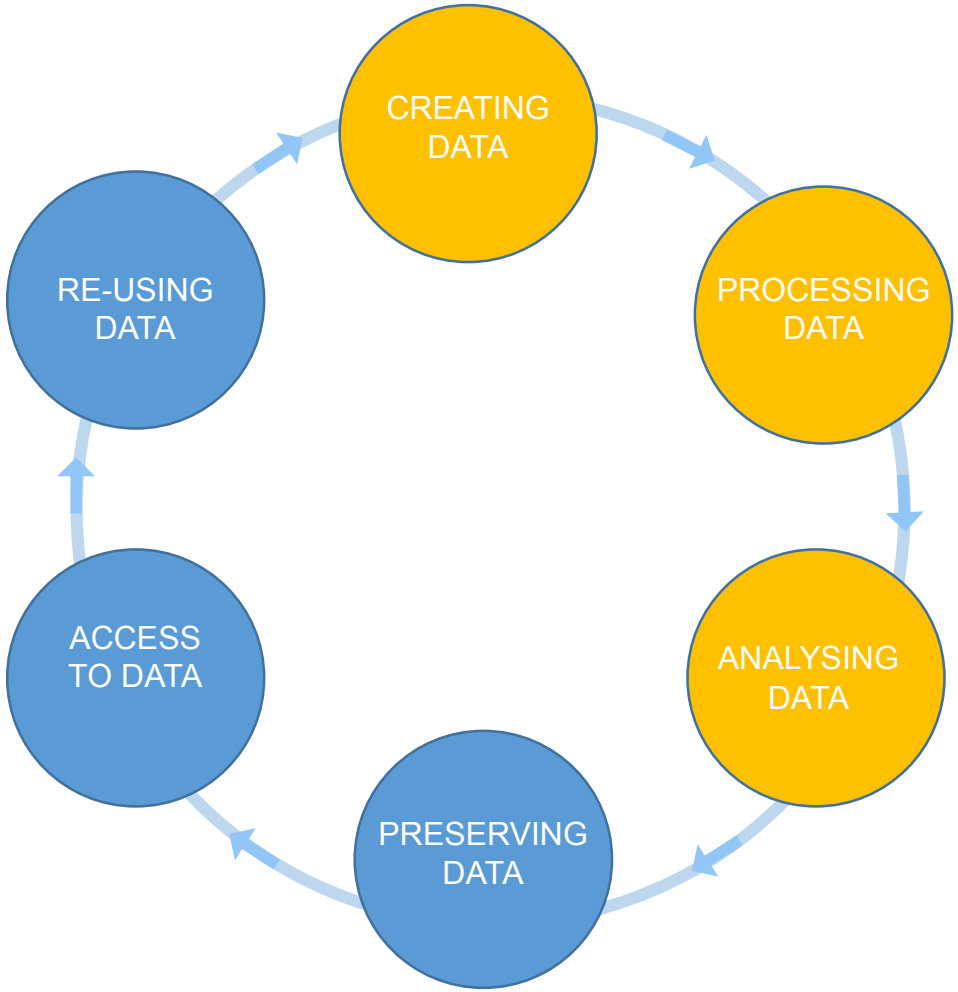
Intermediate Systems

Successful RDM needs Researchers AND Institutions, probably linked through intermediate RDM systems.

Freeman and Millar, “Valuing the Project”, Public Money & Management, 2017

RDM as part of the workflow

Day – to – day, part 1: Structured data



FAIRDOM Platform:

- fine grained metadata,
- logical structure, data relationships

- customizable
- hooks for data processing

- Free, open source.

Troup, E. et al. (2015) Practical evaluation http://hdl.handle.net/1842/12236

FAIRDOM

http://fair-dom.org

Findable
Accessible
Interoperable
Reusable

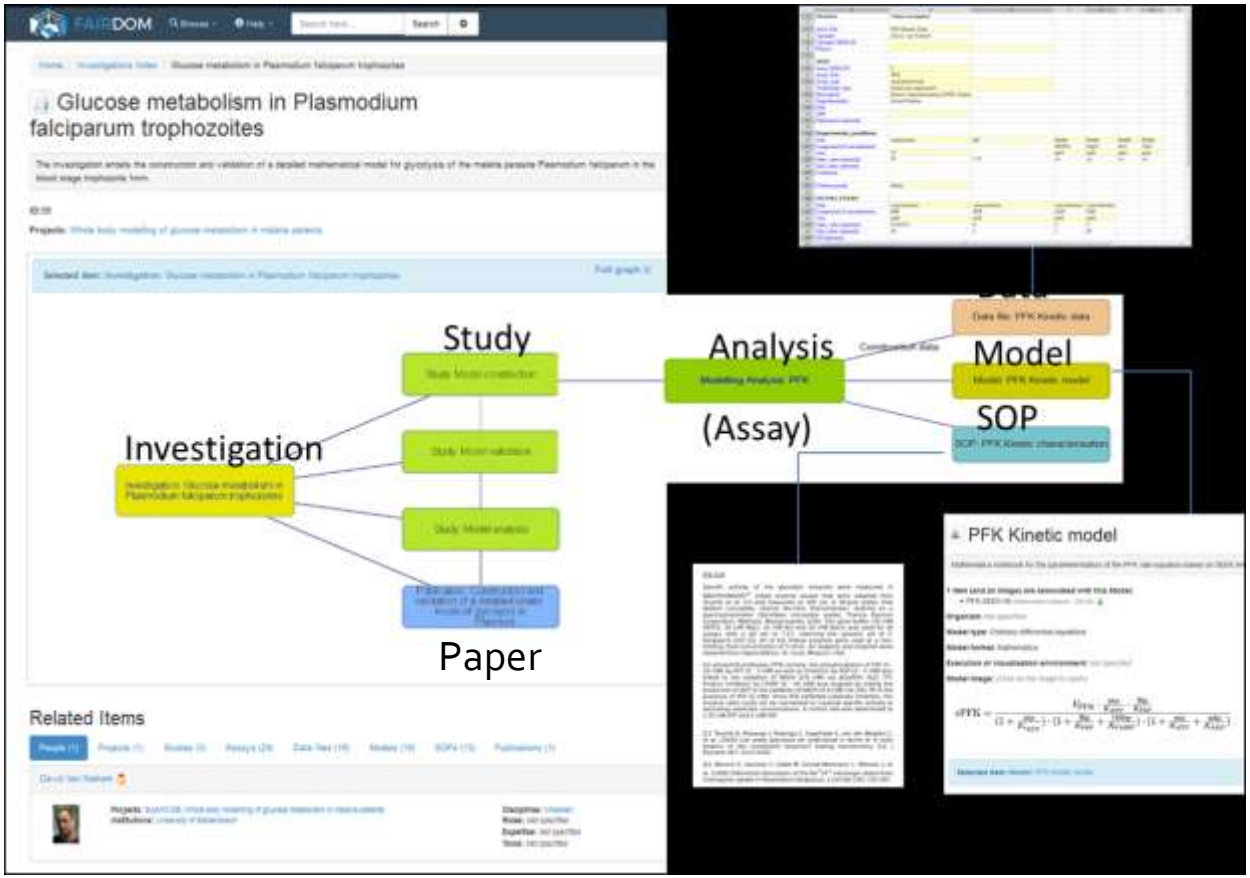
Data
Operations
Models

Data – Models – SOPs- Workflows –

Data and Model Management for all scales of projects

Slide: Prof. Carole Goble, Manchester, 2017.

Organise and Report (snapshot)



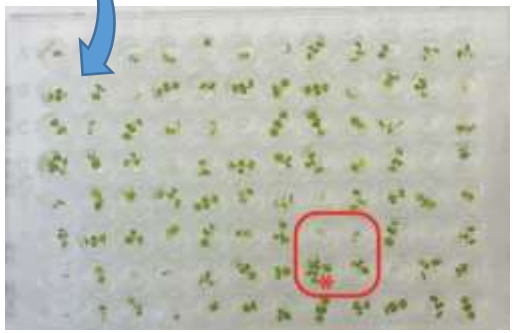
95 investigations
176 studies
345 assays
1398 data files
166 models
214 sops
281 publications
298 presentations
58 events

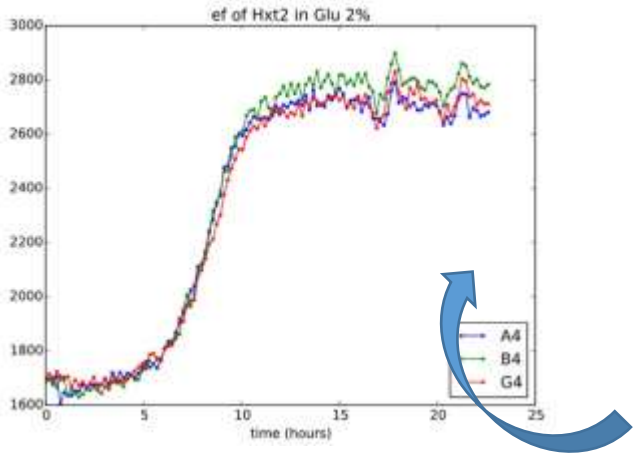
12 projects
127 people
52 organisations
93 ISAs
153 data files
2 models
22 SOPS
3 publications
113 presentations
22 events

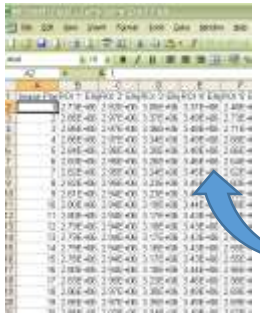
Slide: Prof. Carole Goble, Manchester, 2017.

RDM as part of the workflow

Day – to – day, part 1: FAIRDOM



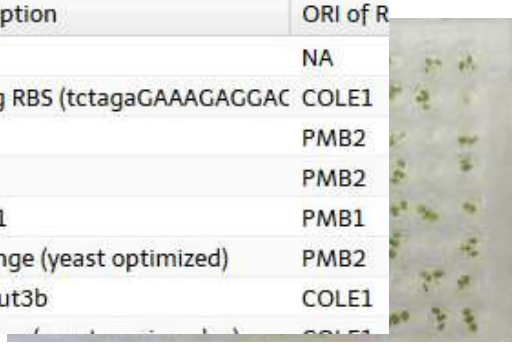


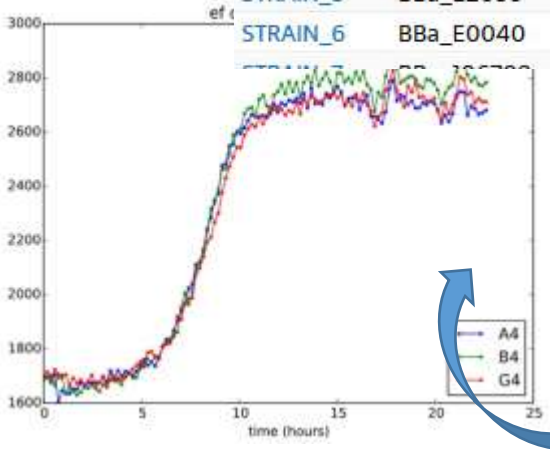


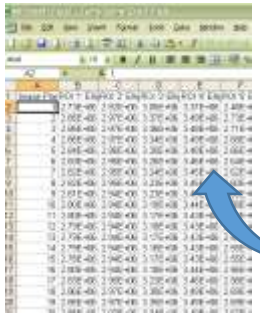
RDM as part of the workflow

Day – to – day, part 1: FAIRDOM

Code	Reporters	Resistance	Backbone	Description	ORI of R
STRAIN_1	E.coli DH10B	NONE	no vectors		NA
STRAIN_10	BBa_B0034	AP	pSB1A2	Strong RBS (tctagaGAAAGAGGAC	COLE1
STRAIN_2	BBa_E0030	CM	pSB1C3	EYFP	PMB2
STRAIN_3	BBa_E0020	CM	pSB1C3	ECFP	PMB2
STRAIN_4	BBa_E1010	CM	pSB1C3	mRFP1	PMB1
STRAIN_5	BBa_E2050	CM	pSB1C3	mOrange (yeast optimized)	PMB2
STRAIN_6	BBa_E0040	AP	pSB1A2	GFPmut3b	COLE1

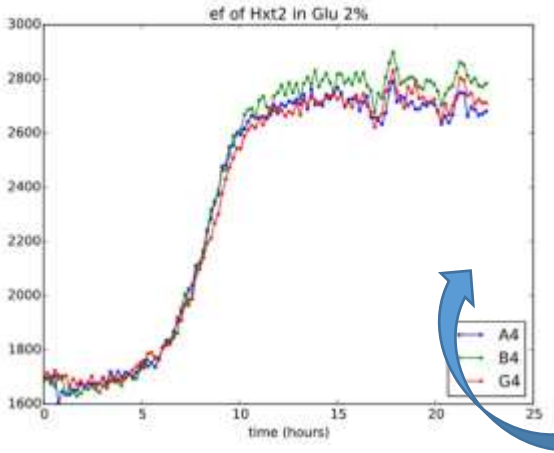




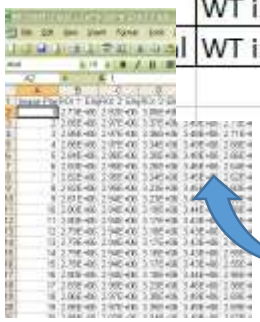


RDM as part of the workflow

Day – to – day, part 1: FAIRDOM

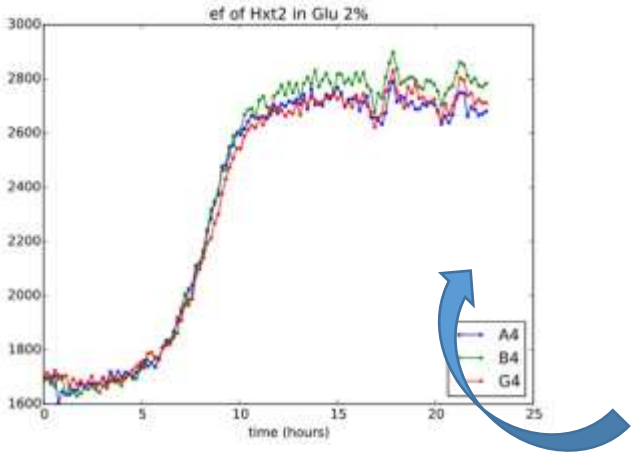


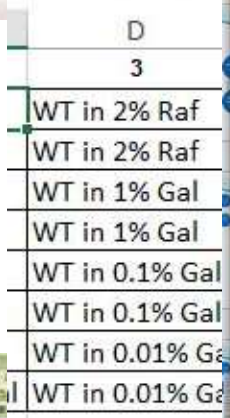
	D	E	F	G
	3	4	5	6
	WT in 2% Raf	WT in 2% Raf	GAL1 in 2% Raf	GAL2 in 2% Raf
	WT in 2% Raf	WT in 2% Raf	GAL1 in 2% Raf	GAL2 in 2% Raf
	WT in 1% Gal	WT in 1% Gal	GAL1 in 1% Gal	GAL2 in 1% Gal
	WT in 1% Gal	WT in 1% Gal	GAL1 in 1% Gal	GAL2 in 1% Gal
	WT in 0.1% Gal	WT in 0.1% Gal	GAL1 in 0.1% Gal	GAL2 in 0.1% Ga
	WT in 0.1% Gal	WT in 0.1% Gal	contaminated	GAL2 in 0.1% Ga
	WT in 0.01% Gal	WT in 0.01% Gal	GAL1 in 0.01% Gal	GAL2 in 0.01% C
l	WT in 0.01% Gal	WT in 0.01% Gal	GAL1 in 0.01% Gal	GAL2 in 0.01% C

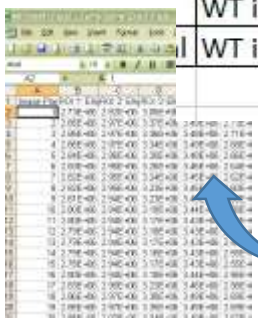


RDM as part of the workflow

Day – to – day, part 1: FAIRDOM



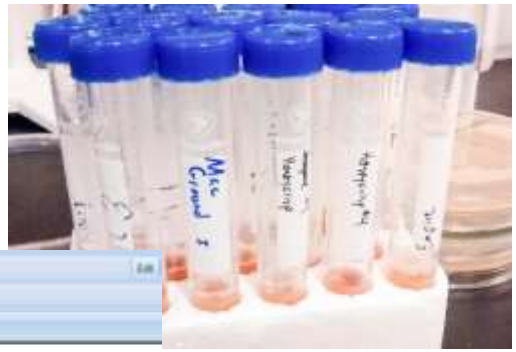


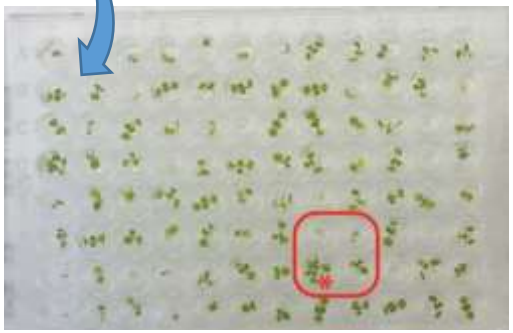


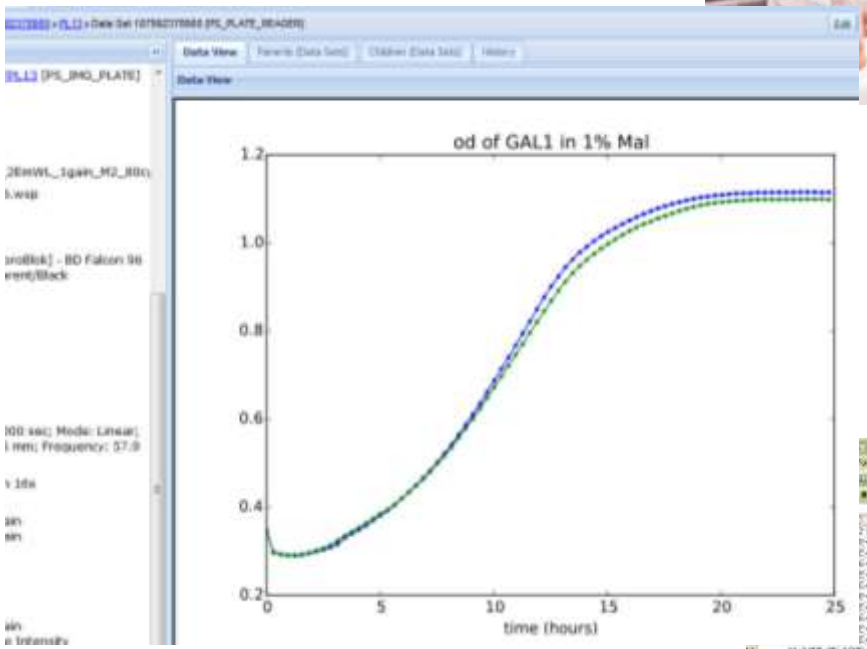
Experiment Type	PS_GROWTH
Registrator	
Registration Date	2015-08-07 14:59:08
Project	/PS_GROUP/GROWTH_RATE
Name	test sugar upload
Description	Testing uploading a sugar file.
Aim	Measure expression of GAL genes in maltose
Authors	Ivan
Submitter	testps
Date	2013-09-12
Measurement date	2012-12-16
Strains	WT GAL1 GAL2 GAL3 GAL7 GAL10 GAL80
Sugars	RAF MAL
Comments	DropBox ERROR: Missing strains in the repository: GAL2,GAL3,GAL7,GAL10,GAL80 Please update the repository and add the correct links to the strains Missing sugars in the repository: MAL

RDM as part of the workflow

Day – to – day, part 1: FAIRDOM

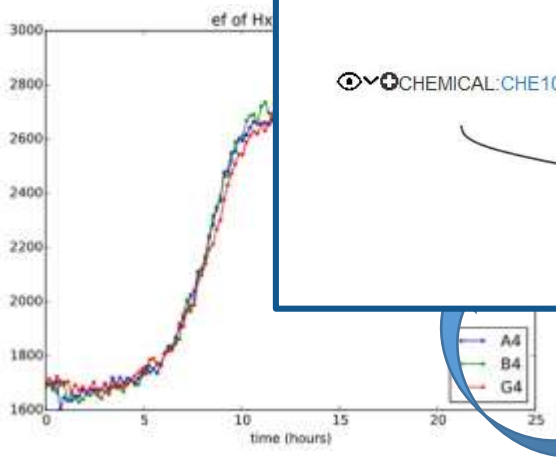


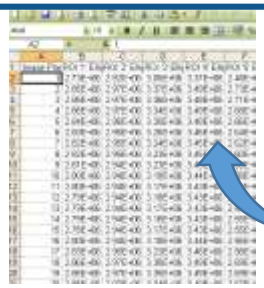




RDM as part of the workflow

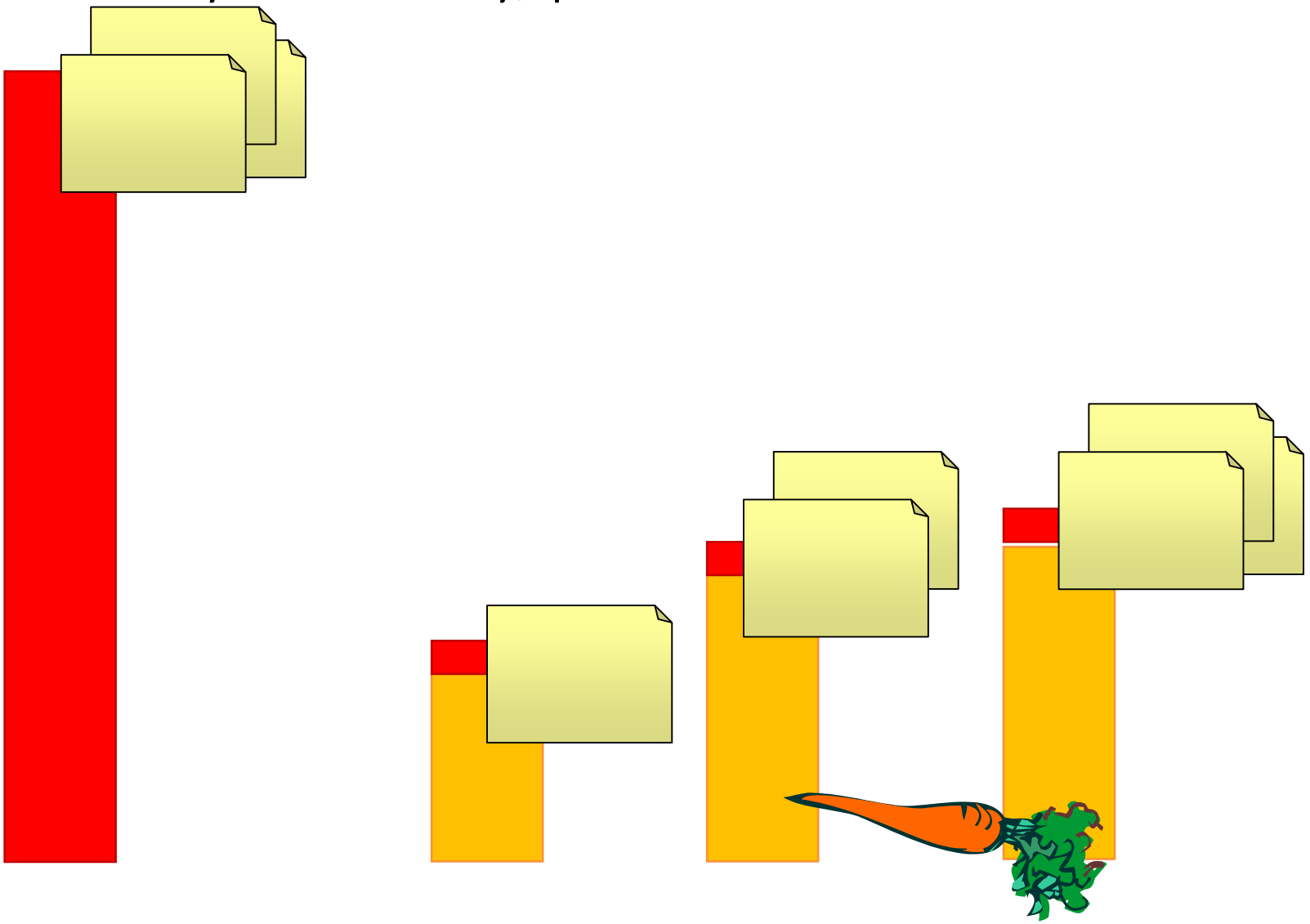
Day – to – day, part 1: FAIRDOM





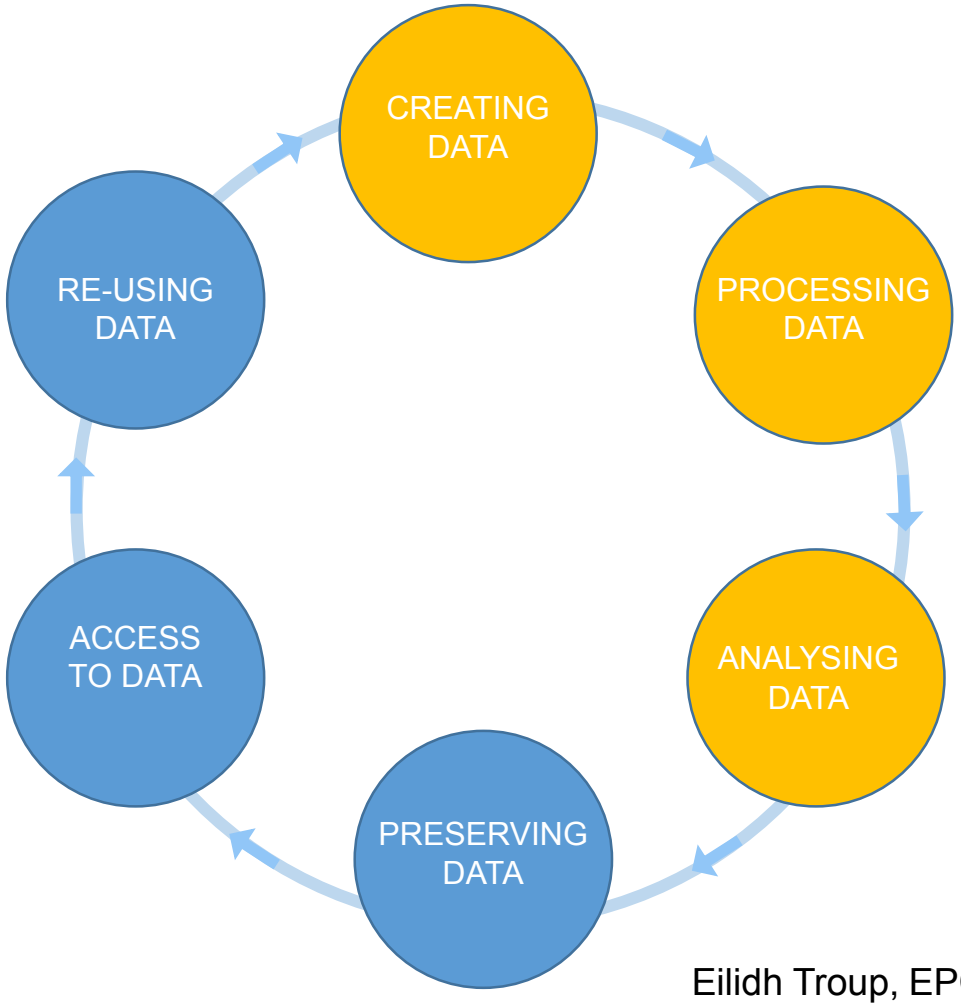
RDM as part of the workflow

Day – to – day, part 1: FAIRDOM



RDM as part of the workflow

Day – to – day, part 2: university wiki as ‘free’, ‘default’ lab notebook



Multiple labs already.

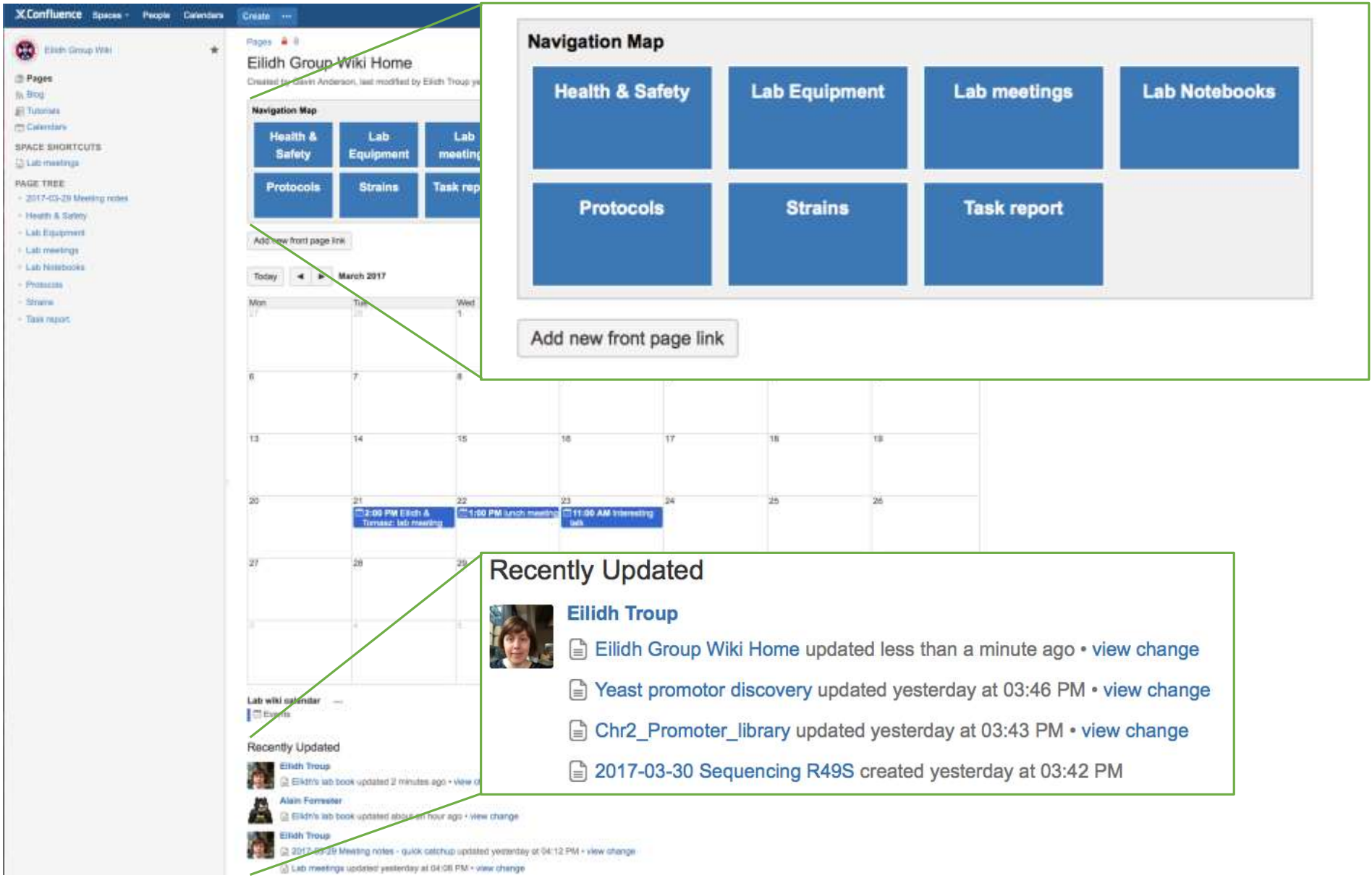
Typical features of ELN:

- Rich Text Formatting
- Embedded images
- File attachments
- Hierarchical structure
- Online, sharing options

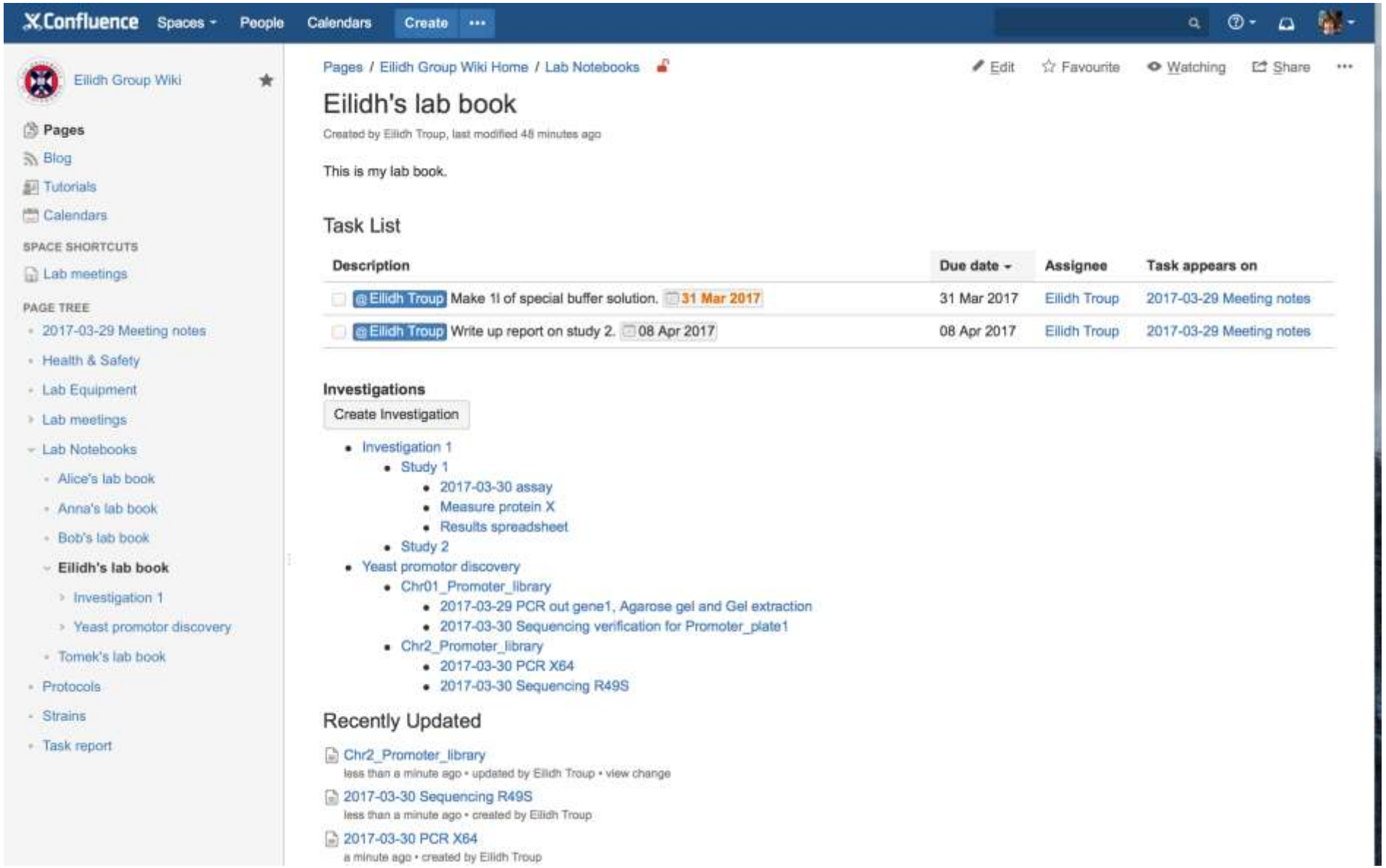
Few carrots but it’s simple

Eilidh Troup, EPCC.
Funded by Wellcome Trust ISSF.

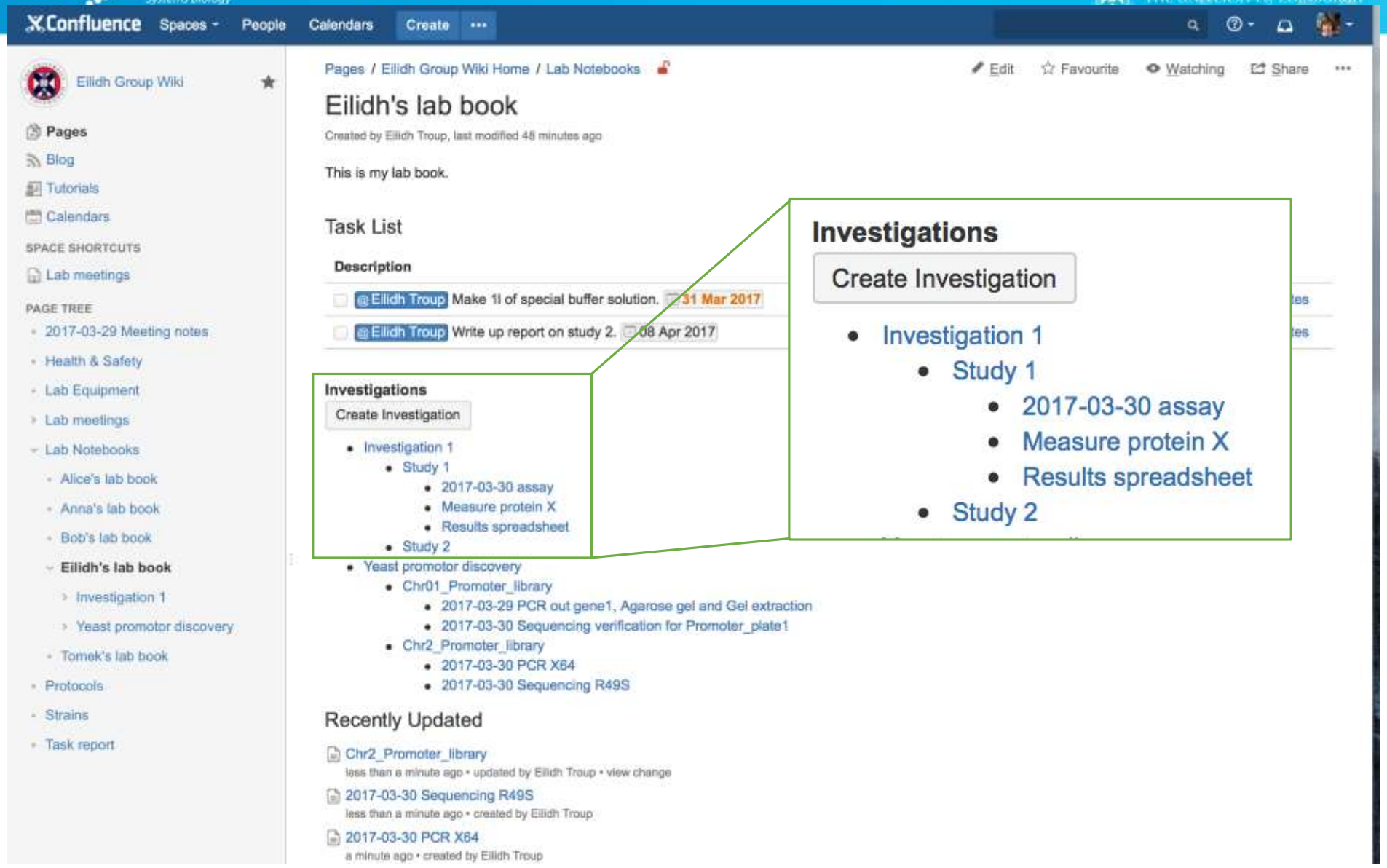
Confluence wiki – group view



Wiki – user view



Wiki – user view



Investigation-Study-**Assay**

Confluence Spaces People Calendars Create ...

Eilidh Group Wiki

Pages
Blog
Tutorials
Calendars

SPACE SHORTCUTS
Lab meetings

PAGE TREE
- 2017-03-29 Meeting notes
- Health & Safety
- Lab Equipment
- Lab meetings
- Lab Notebooks
  - Alice's lab book
  - Anna's lab book
  - Bob's lab book
  - Eilidh's lab book
    - Investigation 1
    - Yeast promotor discovery
      - Chr01_Promoter_library
        - **2017-03-29 PCR out gene1,**
        - 2017-03-30 Sequencing verif
      - Chr2_Promoter_library
  - Tomek's lab book
- Protocols
- Strains
- Task report

Pages /... / Chr01_Promoter_library

Edit

2017-03-29 PCR out gene1, Agarose gel and Gel extraction

Created by Eilidh Troup, last modified 45 minutes ago

After two unsuccessful intends to PCR out the gene1 vector, I decided to redesign the forward primer. After some analysis I realize the original set of primers we

Primers:

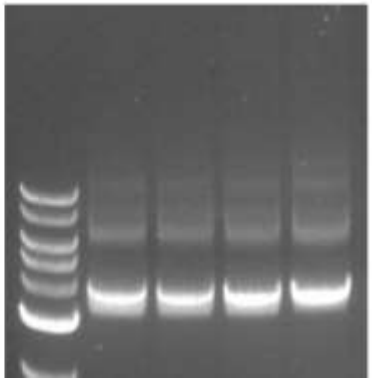
	Sequence 5'-3'	Tm
gene1GGF2	ggtccaatgtggtaaaccatacttagatcgg	60
gene1GGR	gttcgtcacatccttccttcttcttttggtc	60

The primers were diluted to a stock concentration of 100uM and a working solution of 10uM was prepared for cloning.

PCR reaction

Reagent	Volume
200ng DNA	0.42uL
5x Buffer HF	10uL
5mM dNTPS	2uL
10uM Forward Primer	2.5uL
10uM Reverse Primer	2.5uL
Phusion Polymerase	0.5uL
Water	32.08uL

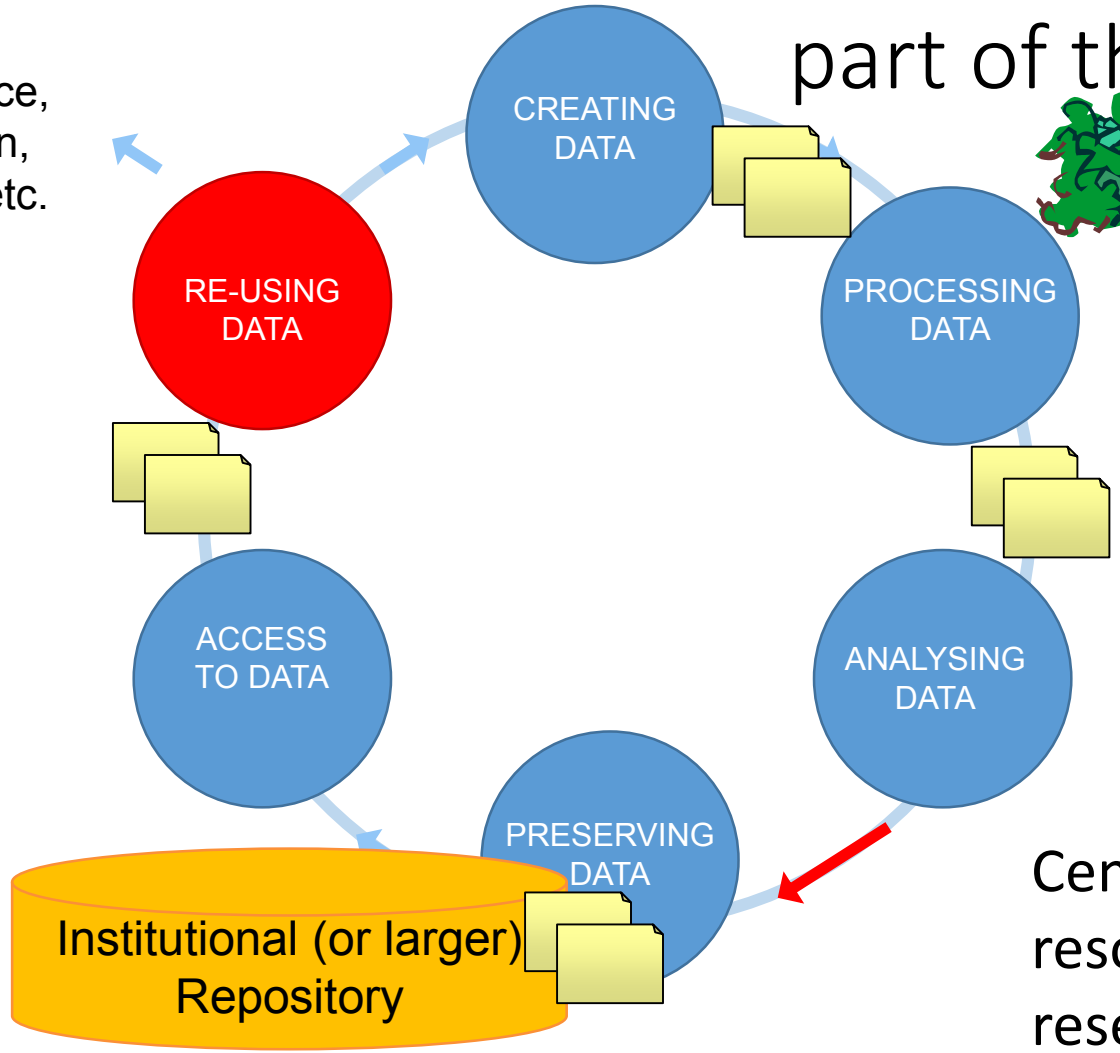
The expected size of the amplicon is 4.1Kb. Comparing the gel band of the 4 replicates with the 1kb ladder the amplicon size is aprox the expected.



Data life-cycle

Data management as part of the workflow

New Science, Publication, Citations, etc.



BioDare2
FAIRDOM
'wikibench'

Centrally-supported resources for "smart" research output